Published online 18 September 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D173–D177***
doi:10.1093/nar/gkm696

piRNABank: a web resource on classified and clustered Piwi-interacting RNAs

S. Sai Lakshmi and Shipra Agrawal*

Institute of Bioinformatics and Applied Biotechnology, Bangalore, India

Received June 21, 2007; Revised July 21, 2007; Accepted August 22, 2007

ABSTRACT

Piwi-interacting RNAs (piRNAs) are expressed in mammalian germline cells and have been identified as key players in germline development. These molecules, typically of length 25–33 nt, associate with Piwi proteins of the Argonaute family to form the Piwi-interacting RNA complex. These small regulatory RNAs have been implicated in spermatogenesis, repression of retrotransposon transposition in germline cells, epigenetic regulation and positive regulation of translation and mRNA stability. piRNABank is a highly user-friendly resource which stores empirically known sequences and other related information on piRNAs reported in human, mouse and rat. The database supports organism and chromosome-wise comprehensive search features including accession numbers, localization on chromosomes, gene name or symbol, sequence homology-based search, clusters and corresponding genes and repeat elements. It also displays each piRNA or piRNA cluster on a graphical genome-wide map (http://pirnabank.ibab.ac.in/).

INTRODUCTION

Eukaryotic gene expression is regulated by a wide variety of RNA species at the transcriptional, post-transcriptional and translational levels. Small non-protein-coding RNAs have gained significant importance due to their widespread occurrence and diverse functions as regulatory molecules, which are essential for cell growth and development in eukaryotes. Micro-RNAs (miRNAs), small interfering RNAs (siRNAs), repeat-associated small interfering RNAs (rasiRNAs) and Piwi-interacting RNAs (piRNAs) are a few well-known small RNAs (1–3). Classification of small regulatory RNAs is based upon their biogenesis, functions and mechanism of action (2). These RNAs associate with the Argonaute group of proteins to perform sequence-specific gene silencing mechanisms, including mRNA degradation, transcriptional gene silencing, translational repression, heterochromatin formation and DNA elimination (4–9). Argonaute proteins act as molecular scaffolds that present the small, guide RNA molecules of RNA silencing to their complementary targets, by forming a ribonucleoprotein complex called RNA-induced silencing complex (RISC) (8,10). Three Argonaute proteins, namely Aub, Piwi and AGO3 (endonucleases) occur in the germline cells and are grouped under the PIWI subfamily of proteins (9,11). Piwi has been shown to be a nuclear protein involved in gene silencing of retrotransposons and controlling their mobility in the male germline (12). It has been reported that knockout mutations in Piwi proteins lead to defects in sperm development (13).

piRNAs are a newly identified class of small regulatory RNAs, abundantly produced in the germline cells of eukaryotes. Piwi-interacting RNA complex (piRC) is a complex of piRNAs with the Piwi protein, extracted and purified from mammalian testes (3–8,14,15). Further, a similar class of small RNAs (rasiRNAs) has been widely studied and recently reported to silence the activity of repeat elements in *Drosophila* (3,11,16,17). Recent updates on piRNAs confirm the role of these small RNAs in regulating transposon mobility and activity in mammals (3,11,15). The length of piRNAs ranges between 19 and 33 nt, most of them fall in the range of 25–33 nt. Like siRNA and miRNA, these RNAs also have a strong preference for the 5′ uridine. Furthermore, these molecules occur in clusters of length 20–100 kb. The piRNA density in these clusters ranges from 40 to 4000 (4–8). Interestingly, these clusters tend to occur on one strand (+/−) or partly on both the strands and are designated as monodirectional clusters and bidirectional clusters, respectively. Bidirectional clusters describe the divergent transcription from the piRNA precursors (1,7). It has been suggested that rasiRNAs found in *Drosophila* are the same as piRNAs identified in the mammalian germline (15). A model of piRNA biogenesis in *Drosophila* has been proposed recently (11,17). However, the mechanism of piRNA production and its mode of action are yet to be elucidated in mammals.

*To whom correspondence should be addressed. Tel: +91 80 2841 0029, 2841 2769; Fax: +91 80 2841 2761; Email: shipra@ibab.ac.in

MOTIVATION FOR piRNABank

Recent reports on piRNAs have revealed the importance of these molecules in the regulation of germline development. The biological significance and the functions of these molecules are currently the subject of intensive study. Research progress has led to the identification of several thousand piRNAs in mammals and a huge amount of data has accumulated in a very short span of time. However, the accessibility to the entire dataset is limited. Currently, there is no piRNA-relevant resource that can fetch data, and annotations, for the user. Illustratively, positional information of piRNAs on the chromosomes is required for studying its association with annotated genomic elements and for identifying the genes being targeted by these regulatory molecules. Since piRNAs have been reported to exist as clusters (4–8), cluster information is very important for researchers. Further, piRNA clusters provide insights into piRNA biogenesis from a single precursor, or two precursors whose transcription is triggered by a common central promoter (1,7). piRNA biology is gaining enormous attention and the need of the hour is to collect and unify the available piRNA data, which would accelerate research in this field. One such bioinformatics resource which provides a collection of all piRNA sequences reported in NCBI nucleotide sequence database is RNAdb, a database on all mammalian non-protein-coding RNAs (18). RNAdb is limited to providing a list of all piRNA sequences and does not allow the user to access any other annotation associated with piRNA data across different chromosomes or clusters.

Considering the biological significance of piRNAs and with the aim of providing easy access to the large and growing volume of data on these molecules, we have developed piRNABank, a repository of all known piRNAs in human, mouse and rat. It would serve as an important tool for molecular biologists studying the biogenesis and regulatory roles of these molecules in mammalian systems and facilitate future research in piRNA-mediated RNA interference. piRNABank is the first known web resource, which provides sequence as well as annotation information on piRNA data from mammals. The piRNA data has been analysed, organized and integrated to develop a highly user-friendly database and analysis system. The web interface enables the user to execute a quick and efficient search on piRNA data. The database can be queried comprehensively through various arguments such as accession number, gene name or symbol, chromosome number, chromosomal position, piRNA clusters in specific chromosomal region(s), total number of clusters in a selected chromosome and clusters with a defined piRNA density. It also facilitates the display of graphical as well as tabular information on the associated genes, repeat elements and corresponding piRNA or piRNA cluster data in a user-selected chromosomal segment. This web analysis system allows searches for piRNA homologues by a simple string matching or BLASTN search. With the availability of the aforementioned features, piRNABank will be an extremely useful resource for computational and experimental biologists working in this and related areas.

DATA PROCUREMENT AND REFINEMENT

piRNA dataset

The large-scale sequencing of piRNAs from rat, mouse and human testes by different experimental groups have yielded a large number of piRNA sequences, which have been reported in the NCBI nucleotide sequence database (4,7,8) and Supplementary Data in the published literature (5,6). The sequences of piRNAs in human, mouse and rat have been downloaded from the NCBI nucleotide sequence database. Apart from the sequences from the NCBI database, experimentally characterized piRNAs (not submitted in the NCBI sequence database) listed in the Supplementary Data of the available literature have also been added to the dataset. Redundancy and repetition in piRNA sequences has been carefully removed at different stages of our analysis to obtain a unique dataset. Exactly matching sequences taken from multiple sources were eliminated while constructing the piRNA dataset. Contig and clone sequences reported in the NCBI sequence database, which were longer than the known length of piRNAs, were also removed.

In order to identify the positions of piRNAs on the chromosomes, whole genome sequences of human (NCBI36), mouse (NCBIM36) and rat (RGSC3.4) were downloaded from Ensembl Genome Browser. WU-BLAST2.0 was installed and configured on the local machine. The parameters used to perform BLAST are as follows: $E = 0.01$; no gaps; W (seed word length for ungapped BLAST, default length is 11 nt for BLASTN) = query sequence length; B (maximum number of database sequences for which alignments will be reported) = 80 000; hspmax (maximum number of ungapped HSPs that will be saved per subject sequence) = 80 000; hspsepSmax (maximum allowed separation along the subject sequence between two HSPs) = 0. Perl and shell scripts were written to parse the BLAST results and obtain the chromosome positional information of piRNAs. The sequences, which did not map to the genome, were not included in the dataset. The dataset has been further refined by the following process. Two or more piRNAs mapping exactly to the same positions on the genome were identified. These piRNAs were compared with each other and found to have the same nucleotide sequence, with one or more extra bases either at 5′ or 3′ end. The longer sequence was retained in the dataset. Currently, piRNABank harbours 23 439 human, 39 986 mouse and 38 549 rat unique piRNA sequences, which are mapping to unique or multiple loci on the corresponding genome. The entire dataset maps to 667 944, 1 399 813 and 1 269 304 positions on human, mouse and rat genomes, respectively. Exact number of sequences involved in the generation of piRNA dataset at different levels has been summarized in Table 1.

DATABASE STRUCTURE AND CONTENT

Data have been stored in relational tables in a MySQL database. A specific naming convention has been used to uniquely identify each piRNA sequence in piRNABank.

Table 1. piRNA dataset: statistics indicating the numbers of piRNAs at different levels of data collection, organization and generation of the piRNA dataset (as available in piRNABank) for the three organisms human, mouse and rat

Organisms	Human	Mouse	Rat
Total number of sequences downloaded from NCBI	11 147 151	6 943 119	1 334 465
Sequences from the literature (Supplementary Data)	1005	4045	–
Number of piRNAs after extensive removal of redundancy (contigs, clones, repetitive sequences)	32 194	72 878	62 713
Sequences which mapped with the genome	32 194	58 320	55 442
Final piRNAs in piRNABank after further removal of redundancy	23 439	39 986	38 549
Unique mappers on the genome	19 260	35 796	34 243
Total positions mapped on the genome (including sequences mapping to single and multiple loci)	667 944	1 399 813	1 269 304

Human sequences are named from hsa_piR_000001 to hsa_piR_023439. Similarly, prefixes of 'mmu' and 'rno' have been used for naming the mouse and rat piRNAs, respectively. The structure of the tables is identical for human, mouse and rat piRNAs. Supplementary Figure 1 gives a detailed schematic of the organization, flow and structure of data in piRNABank.

piRNA map

To derive piRNA association with the annotated genomic elements such as genes and repeats, the chromosome-specific repeat element table (RepBase Update version 9.11—RM database version 200501112) has been downloaded from UCSC Genome Browser for all three organisms. Gene information has been obtained from NCBI Entrez Gene database. MySQL tables were created in RDBMS for storing the gene and repeat annotation data. These tables have been related to the piRNA data tables. A set of in-house CGI-Perl programs have been used to identify the piRNAs, associated repeats and genes in a selected chromosomal region, which is graphically displayed as a map.

piRNA clusters

Highly stringent criteria have been used to identify the clusters. Based on already defined rules for identifying clusters, piRNA density in a region of the chromosome has been used as threshold cutoff (4,5,7). The rules used in data clustering are given as follows:

(i) Each chromosome is scanned using a 20 kb sliding window, with 1 kb increments.
(ii) The window having more than one specific threshold cutoff of uniquely mapping piRNAs is extracted. Threshold values have been designated based on the piRNA density for each organism.
(iii) All such windows satisfying the threshold are merged together and every 1 kb of the probable cluster is checked for the presence of at least 2 piRNAs.
(iv) Exact cluster boundaries are found by trimming the right- and left-side boundaries of the probable cluster by 100 bases towards the centre of the cluster.

This clustering algorithm has led to the identification of 89 piRNA clusters in human, 111 clusters in mouse and 189 clusters in rat. Information on piRNA cluster positions on the chromosome and their strand specificity has been stored in separate tables in the MySQL database.

SOFTWARE AND IMPLEMENTATION

The interface layer of piRNABank has been developed using HTML, DHTML and JavaScript. piRNA data and information on the associated genomic elements have been stored in MySQL relational database tables. The application layer between the web interface and the back-end relational tables has been implemented using CGI-Perl. All computational programs for the collection, sorting and redundancy removal of the data and the genome mapping and clustering of piRNAs have been written in Perl and Linux shell scripting languages.

piRNABank primarily processes the user query through simple search and advanced search options, which in turn retrieves information from the relational database tables, formats the result and displays it on the web interface. Sequence and cluster information have been stored in the form of a flat file database, which is used for data downloading.

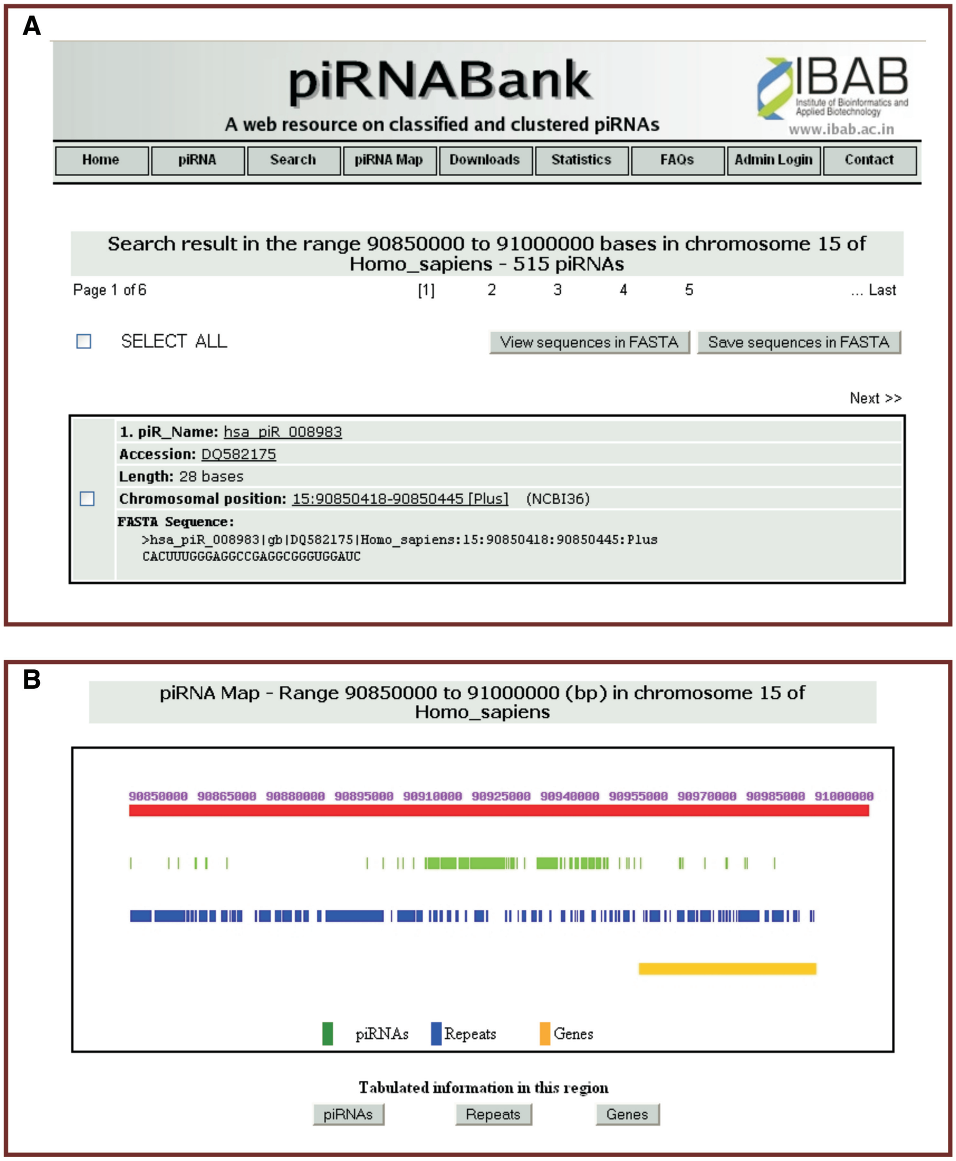
DATA ACCESS

Data stored in piRNABank can be accessed in the following ways:

(i) Search options in piRNABank: piRNABank can be queried to obtain piRNA information in many ways. In order to facilitate this, simple and advanced search options have been provided in the 'Search' section.

   A 'simple search' can be performed using the following parameters:

   (a) piRNABank or NCBI accession numbers: the user can enter the piRNABank or NCBI accession numbers to obtain piRNA sequence information.
   (b) Gene name or symbol: the user can select the organism and specify the gene name or gene symbol to view all piRNAs overlapping with specified gene(s).
   (c) Chromosome number and/or by specifying the genomic position: the user can select the organism name to view all piRNAs of the selected organism. The chromosome number can also be chosen to obtain the relevant piRNA data of the


A
piRNABank
A web resource on classified and clustered piRNAs
IBAB
Institute of Bioinformatics and Applied Biotechnology
www.ibab.ac.in
Home
piRNA
Search
piRNA Map
Downloads
Statistics
FAQs
Admin Login
Contact

Search result in the range 90850000 to 91000000 bases in chromosome 15 of Homo_sapiens - 515 piRNAs

Page 1 of 6 [1] 2 3 4 5 ... Last

□ SELECT ALL

□ **1. piR_Name:** hsa_piR_008983
Accession: DQ582175
Length: 28 bases
Chromosomal position: 15:90850418-90850445 [Plus] (NCBI36)
FASTA Sequence:

```
>hsa_piR_008983|gb|DQ582175|Homo_sapiens:15:90850418:90850445:Plus
CACUUUGGGAGGCCGAGGCGGGUGGAUC
```

Figure 1. Screen shot of search result from piRNABank. (**A**) An output page of a chromosome-based search, showing piRNAs in the region from 90850000 to 91000000 in human chromosome 15. The hit count of 515 piRNAs is displayed across six pages on the web interface. A sample piRNA entry on the result page has been shown in the figure. (**B**) piRNA map of the selected region on human chromosome 15; it shows piRNAs, associated retro elements and genes on a chromosomal scale.

selected organism. Additionally, the user can enter the genomic region (in bp) to view all piRNAs in the selected region of a particular chromosome. Figure 1a illustrates the result of chromosome-based search.

The 'advanced search' page allows searching piRNABank with the following options:

(a) Search piRNA clusters: an organism can be chosen to view all piRNA clusters. The user can also select the chromosomal number and position in the selected organism to obtain cluster information. Alternatively, clusters having a specific number or range of piRNAs can also be queried using this search option.

(b) Search homologous piRNAs: users can enter query sequences to identify homology with the piRNA sequences in the dataset. Furthermore, piRNA homologues can be identified by string searching, wherein short query sequences can be searched for matches against the database sequences. Additionally, a BLASTN search allows the user to identify exactly matching sequences stored in piRNABank. Users can specify the e-value cutoff and the maximum number of hits to be reported.

The results are formatted and displayed in the form of tables in the web interface. It provides extensive information on each piRNA, including accession number, sequence length, chromosome number, genomic start and end position, strand orientation and FASTA sequence. Furthermore, the accession number, literature reference and the chromosomal position of each entry has been externally linked with the NCBI nucleotide sequence database, PubMed, NCBI Map viewer and Ensembl genome browser.

(ii) The 'piRNA Map' is an extremely useful feature for visualization of piRNAs and the associated genes and repeats on a genome-wide map. Users can select the organism, chromosome number and the region on the chromosome to view the piRNAs and other annotations. All information is made available as tables as well. Figure 1b shows a sample piRNA map generated by piRNABank.

(iii) Batch download options of piRNA sequences and clusters on specific chromosomes in human, mouse and rat have been provided in the 'Downloads' section. The entire piRNA data on each organism can also be downloaded in FASTA sequence format. Users can extract and download cluster-specific piRNA information for analysis of piRNA precursors. Downloads section also provides piRNA data mapping to previous as well as current genome assemblies.

FUTURE WORK

piRNABank is proposed as a central repository on piRNAs. The resource will be updated constantly with further enhanced features. We also intend to add tools on structural and sequence motif prediction. The piRNA information on *Drosophila* and other organisms will be included in the database as and when data is reported.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors wish to record their gratitude to Prof. N. Yathindra for suggesting this problem and all the support. We thank Dr Gayatri Saberwal for her help in improving the manuscript and Ms J. Janani, project trainee at IBAB for technical assistance. Our sincere thanks are due to the two anonymous reviewers for their invaluable suggestions in significantly improving the database and manuscript. The Open Access publication charges for this article were waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Kim,V.N. (2006) Small RNAs just got bigger: Piwi-interacting RNAs (piRNAs) in mammalian testes. *Genes Dev*., **20**, 1993–1997.
2. Tolia,N.H. and Joshua-Tor,L. (2007) Slicer and the Argonautes. *Nat. Chem. Biol*., **3**, 36–43.
3. O'Donnell,K. and Boeke,J. (2007) Mighty Piwis defend the germline against genome intruders. *Cell*, **129**, 37–44.
4. Girard,A., Sachidanandam,R., Hannon,G.J. and Carmell,M.A. (2006) A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature*, **442**, 199–202.
5. Aravin,A., Gaidatzis,D., Pfeffer,S., Lagos-Quintana,M., Landgraf,P., Iovino,N., Morris,P., Brownstein,M.J., Miyagawa,S.K. *et al*. (2006) A novel class of small RNAs bind to MILI protein in mouse testes. *Nature*, **442**, 203–207.
6. Grivna,S.T., Beyret,E., Wang,Z. and Lin,H. (2006) A novel class of small RNAs in mouse spermatogenic cells. *Genes Dev*., **20**, 1709–1714.
7. Lau,N.C., Seto,A.G., Kim,J., Miyagawa,S.K., Nakano,T., Bartel,D.P. and Kingston,R.E. (2006) Characterization of the piRNA complex from rat testes. *Science*, **313**, 363–367.
8. Watanabe,T., Takeda,A., Tsukiyama,T., Mise,K., Okuno,T., Sasaki,H., Minami,N. and Imai,H. (2006) Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. *Genes Dev*., **20**, 1732–1743.
9. Carmell,M.A., Girard,A., van de Kant,H.J., Bourc'his,D., Bestor,T.H., de Rooij,D.G. and Hannon,G.J. (2007) MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. *Dev. Cell*, **12**, 503–514.
10. Paroo,Z., Liu,Q. and Wang,X. (2007) Biochemical mechanisms of the RNA-induced silencing complex. *Cell Res*., **17**, 187–194.
11. Gunawardane,L.S., Saito,K., Nishida,K.M., Miyoshi,K., Kawamura,Y., Nagami,T., Siomi,H. and Siomi,M.C. (2007) A slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in *Drosophila*. *Science*, **315**, 1587–1590.
12. Saito,K., Nishida,K.M., Mori,T., Kawamura,Y., Miyoshi,K., Nagami,T., Siomi,H. and Siomi,M.C. (2006) Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the *Drosophila* genome. *Genes Dev*., **20**, 2214–2222.
13. Parker,J.S. and Barford,D. (2006) Argonaute: a scaffold for the function of short regulatory RNAs. *Trends Biochem. Sci*., **31**, 622–630.
14. Carthew,R.W. (2006) A new RNA dimension to genome control. *Science*, **313**, 305–306.
15. Lin,H. (2007) piRNAs in the germ line. *Science*, **316**, 397.
16. Pélisson,A., Sarot,E., Payen-Groschêne,G. and Bucheton,A. (2007) A novel repeat-associated small interfering RNA-mediated silencing pathway downregulates complementary sense gypsy transcripts in somatic cells of the *Drosophila* ovary. *J. Virol*., **81**, 1951–1960.
17. Brennecke,J., Aravin,A.A., Stark,A., Dus,M., Kellis,M., Sachidanandam,R. and Hannon,G.J. (2007) Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell*, **128**, 1089–1103.
18. Pang,K.C., Stephen,S., Dinger,M.E., Engstrom,P.G., Lenhard,B. and Mattick,J.S. (2007) RNAdb 2.0 – an expanded database of mammalian non-coding RNAs. *Nucleic Acids Res*., **35**, D178–D182.